# Nos bastidores da Gramateca: uma série de serviços

Alberto Simões[1] and Diana Santos[2]

[1] Linguateca/Universidade do Minho
ambs@ilch.uminho.pt
[2] Linguateca/Universidade de Oslo
d.s.m.santos@ilos.uio.no

**Resumo** Apresentamos aqui um conjunto de serviços que permitem investigar, comparar e criar exercícios e materiais didáticos, e que constituem parte da infraestrutura necessária para a Gramateca, um ambiente para estudar a gramática da língua portuguesa com base em corpos.

**Keywords:** corpos, interrogação, uso, gramática, Gramateca, AC/DC, serviços

## 1 Enquadramento

Textos contendo material anotado constituem a matéria prima da maior parte da investigação linguística moderna. Contudo, analogamente ao problema de excesso de informação na internete, que exige ferramentas especializadas como motores de busca e vários serviços especializados, também a procura e visualização de corpos linguísticos requer ferramentas que auxiliem o pesquisador.

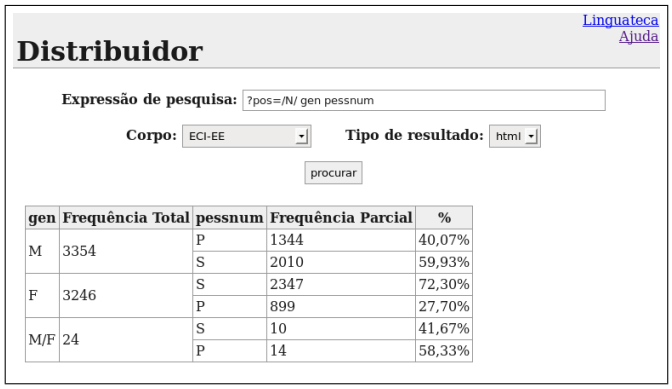
Na Linguateca damos acesso à maior quantidade de informação em forma de corpos sobre a língua portuguesa que existe publicamente, graças ao grande número de projetos e pesquisadores que nos deram autorização para (re)disponibilizar os seus corpos e a respetiva anotação, através do projeto AC/DC [14], descrito em vários artigos recentes [2,10,11]. O AC/DC, ficamos felizes por dizê-lo, não só está em franca expansão como continua a aumentar a abrangência e a quantidade e qualidade da própria informação associada aos corpos. É o AC/DC que constitui a infraestrutura para a Gramateca[3] [12], um projeto recentemente lançado para desenvolver estudos gramaticais do português com base em corpos, dando acesso às fontes e às interpretações usadas.

É importante indicar que todos estes serviços são desenvolvidos sobre o sistema Open Corpus Workbench (Open-CWB) [3], que é, em nossa opinião, o melhor sistema de gestão e indexação de corpos[4]. As ferramentas que têm vindo a ser desenvolvidas mais recentemente são desenvolvidas na linguagem Perl, suportadas pelo módulo `CWB::CQP::More`[5], que disponibiliza uma API de alto nível sobre os módulos originais do Open-CWB.

[3] `http://www.linguateca.pt/Gramateca/`

[4] Embora os autores do Open-CWB também distribuam um sistema para interface na rede, o *CQPWeb*, a Linguateca desenvolveu os seus serviços de raiz porque o início do AC/DC foi anterior a este "novo" produto.

[5] `https://metacpan.org/release/CWB-CQP-More`




| gen | Frequência Total | pessnum | Frequência Parcial | % |
|---|---|---|---|---|
| M | 3354 | P | 1344 | 40,07% |
| | | S | 2010 | 59,93% |
| F | 3246 | S | 2347 | 72,30% |
| | | P | 899 | 27,70% |
| M/F | 24 | S | 10 | 41,67% |
| | | P | 14 | 58,33% |

**Figura 1.** Uso do distribuidor: Distribuição do número (singular ou plural) por género ECI-EE.

Existe, pois, uma ligação direta entre os corpos interrogáveis pelo AC/DC na Gramateca, e os serviços que descreveremos neste texto. Contudo, a parte computacional dos serviços é simples e esperamos que possa ser portada para outros projetos que usem também o Open-CWB sem grandes dificuldades[6]. Por outro lado, o AC/DC – e a Gramateca – estão abertos a todos quantos queiram participar e/ou partilhar os seus dados.

Embora a interface básica, que permite concordâncias e distribuição, seja a mais antiga (15 anos de existência) e a mais utilizada, ao longo do tempo fomos criando outras formas de interagir com os corpos do AC/DC que permitissem fazer certo tipo de pesquisas mais facilmente[7]. Por isso, desde muito cedo existe um serviço simples para prestar informação sobre a frequência de palavras e lemas, o *Ordenador*, e desde há alguns anos temos gizado outros serviços mais especializados, que passamos a descrever aqui, excetuando o *VARRA*, por já ter sido extensamente descrito [7,5].

Neste artigo damos exemplos e explicamos a motivação, mas não pretendemos naturalmente apresentar um manual de utilizador do ambiente do Open-CWB e das suas múltiplas funcionalidades. Para tal o leitor é remetido à extensa documentação do CWB e da Linguateca a esse respeito, contendo tutoriais, listas, exemplos, resposta a perguntas, material didático e artigos de divulgação.[8]

## 2 Distribuidor

O primeiro serviço que apresentamos aqui, de nome *Distribuidor*[9], é uma extensão da distribuição do AC/DC que permite fazer distribuições encaixadas (e que, portanto, é

[6] Recentemente a Linguateca iniciou o processo de disponibilização do código fonte de algumas das suas ferramentas, usando para isso o grupo *Linguateca* no *GitHub*: `https://github.com/linguateca`

[7] Uma preocupação semelhante e uma evolução análoga foram sucedendo com corpos paralelos, como se pode apreciar em [13] e será brevemente mencionada na secção 4.

[8] Veja-se a página intitulada "PJR e exemplos", na área do AC/DC, `http://www.linguateca.pt/ACDC/` para uma visão conjunta dessa documentação.

[9] Acessível de `http://www.linguateca.pt/Distribuidor`.

semelhante em funcionalidade ao comando `cwb-scan-corpus` da bancada Open-CWB — e usa a mesma sintaxe deste).

Alguns exemplos:

Podemos estar interessados em ver de uma forma condensada os substantivos num corpo, em particular qual a sua distribuição por género e número. Ou seja, quantos destes no masculino, e quantos no feminino? Até aqui, isto é possível de fazer na interface padrão do AC/DC com a expressão "`[pos="N"]`" e escolhendo como opção a *Distribuição de género morfológico*;

Mas, e se nos perguntarmos se o plural é aplicado mais vezes para o masculino ou para o feminino? Visto que precisamos da repartição por mais de uma categoria (neste caso, género e número), temos de usar o distribuidor. Na Figura 1 vemos o resultado destas perguntas usando o Distribuidor.

## 3 Comparador

O *Comparador*[10] é parecido com o *Distribuidor*, mas é mais poderoso, porque foi pensado para comparar duas situações que não têm necessariamente que estar incluídas num mesmo corpo, ou que não possam ser distinguidas apenas por um valor diferente de um atributo. Como limitação, repare-se que são apenas duas situações, ou seja, há duas procuras ou pesquisas cujo resultado é mostrado em paralelo, dependendo totalmente do investigador a pertinência das mesmas.

Vejamos alguns exemplos:

Quais as palavras mais correntes associadas com a cor verde ou com a cor azul? Uma forma muito simples (que tem de ser, naturalmente, refinada) encontra-se na parte superior da Figura 2, extraída de [16]. Outro exemplo de uso do Comparador, agora juntando os resultados numa mesma tabela, é a distribuição da menção ao Verão por semana nas duas variantes do CHAVE, na parte inferior da mesma figura.

Após exportado em formato de texto separado por tabuladores (tsv), permite por sua vez a sua visualização usando outras ferramentas. Na figura 3 mostramos como, através do R [8], se pode obter uma representação gráfica da distribuição — um exemplo da interação frutífera entre a investigação e o ensino, visto que é um dos exemplos usados por nós para ensinarmos o conceito estatístico de correlação.

## 4 Ensinador "paralelo"

Já apresentámos o *Ensinador* [17] em 2011, e remetemos o leitor interessado para essa referência. Mas queremos mencionar que houve melhorias significativas desde essa altura, como a possibilidade de ter mais do que um "espaço para resposta" nos exercícios, ou seja, ser possível remover várias palavras, além de ser possível aceitar contextos maiores. Isto implicou uma atualização da sintaxe permitida pelo Ensinador, como o exemplo da Figura 4 demonstra.

[10] Disponível de `http://www.linguateca.pt/Comparador`.

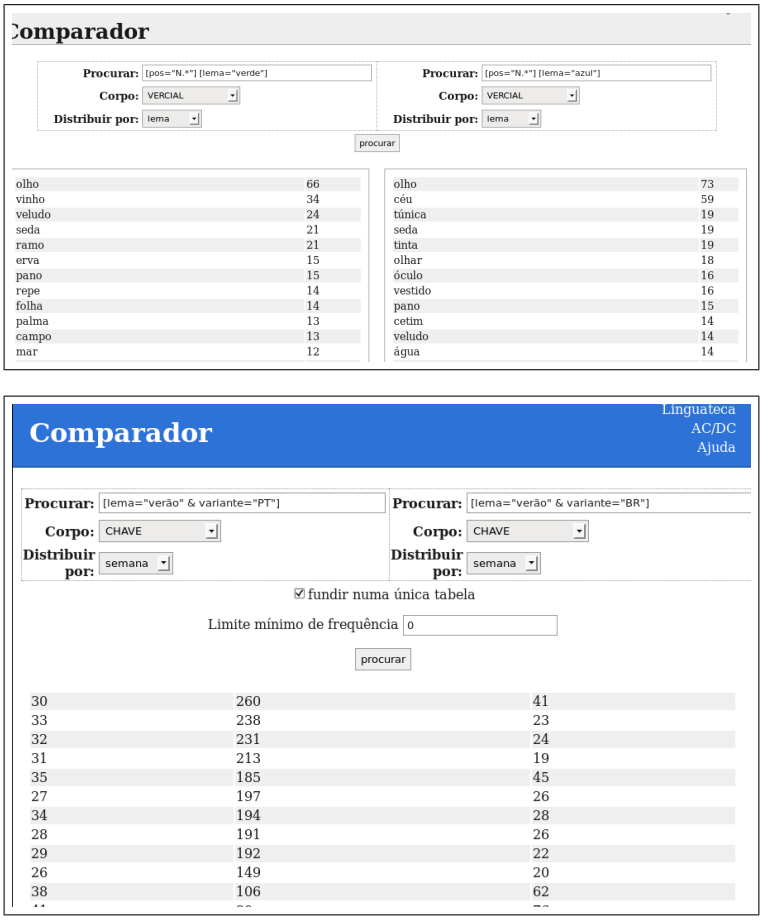



**Figura 2.** Comparador (versão anterior) usado para analisar a distribuição dos substantivos coocorrendo com duas cores diferentes; Comparador analisando a distribuição de *verão* por semana.

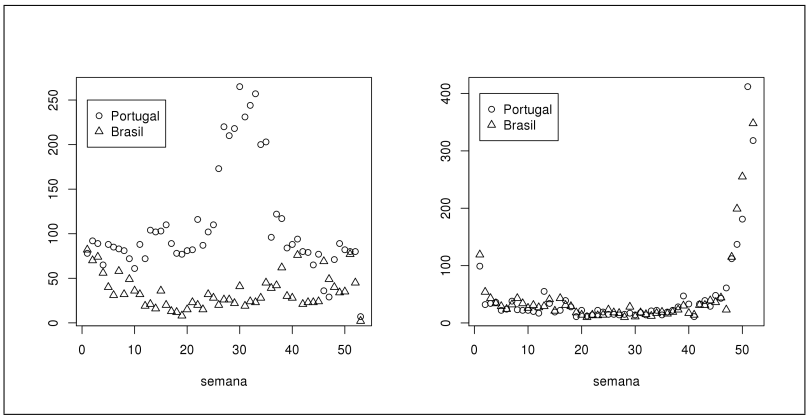



**Figura 3.** A distribuicao de *verão* por semana nas duas variantes do CHAVE (à esquerda), comparada com a distribuição de *Natal* (à direita).

Esta funcionalidade era necessária por duas razões: por um lado, para ter mais liberdade na determinação das concordâncias que depois são objeto de escolha pelo professor, mas sem que isso se torne visível (e desagradável à vista) na solução: claramente não precisam de aparecer em negrito); por outro lado, essencial que sejam mostradas no enunciado para que o texto faça sentido.

Mas além desta melhoria, que é considerável em termos do aumento da riqueza dos enunciados, adicionámos uma nova vertente ao Ensinador, usando corpos paralelos, que permite fazer exercícios bilingues, como mostramos na figura 5. É discutível se devemos considerar este um novo serviço, ou apenas uma extensão do Ensinador, visto que, se por um lado, nasceu como uma extensão, é para aplicar a objeto distintos e terá de ter funcionalidades que só fazem sentido quando estamos na presença de duas línguas.

Note-se que, além dos corpos criados pela Linguateca, tal como o COMPARA, o CorTrad, o PoNTE e o PANTERA[11], existe uma grande quantidade de corpos candidatos a serem utilizados neste Ensinador Paralelo, provenientes do Projeto Per-Fide [1]. Isto fez aliás com que desde o início tenhamos dado mais importância à portabilidade e independência no caso deste serviço do que no dos outros.

## 5 Rêve

Finalmente, o último sistema — e o mais relacionado com um dos traços mais originais da Gramateca , veja-se [15] — é o *Rêve* (corruptela de Revê), que permite rever e reanotar materais linguísticos variados, e cujo objetivo é não só tornar estes acessíveis como permitir o seu fácil manuseamento e a medição de diferenças concetuais.

[11] Respetivamente, `http://www.linguateca.pt/COMPARA`, `http://www.linguateca.pt/CorTrad`, `http://www.linguateca.pt/PoNTE/` e `http://www.linguateca.pt/PANTERA`.

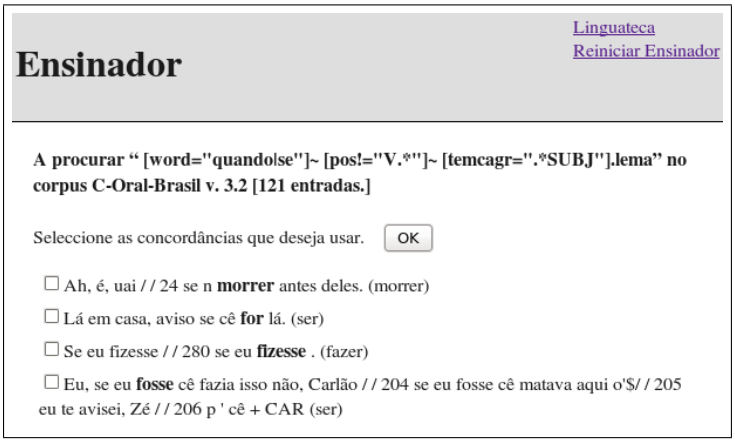



**Figura 4.** Enunciado com sintaxe complicada: procura de formas do presente do conjuntivo a seguir a *quando* ou *se*, mas não esconde a conjunção nem a palavra seguinte.

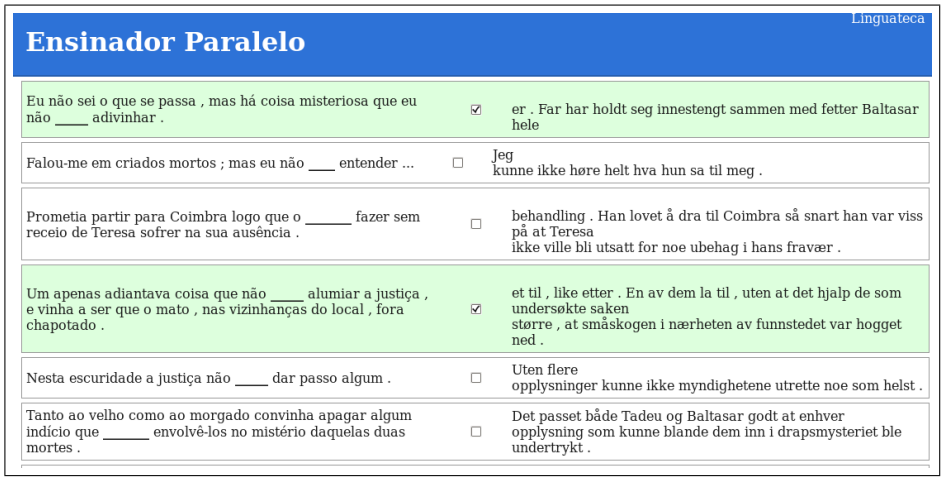



**Figura 5.** Exemplo de concordâncias paralelas para serem utilizadas na discussão de problemas contrastivos ou no ensino da tradução.

Embora o *Rêve* ainda esteja neste momento em desenvolvimento, já permite a inspeção e nova classificação de conjuntos de dados, assim como a sua exportação num formato separado por tabuladores.

Na figura 6, podemos ver uma primeira sua aplicação ao estudo do corpo humano na língua portuguesa, no projeto *Esqueleto* – mais informação sobre o próprio assunto e a complexidade do esquema de anotação podem ser encontrados em [6,4].

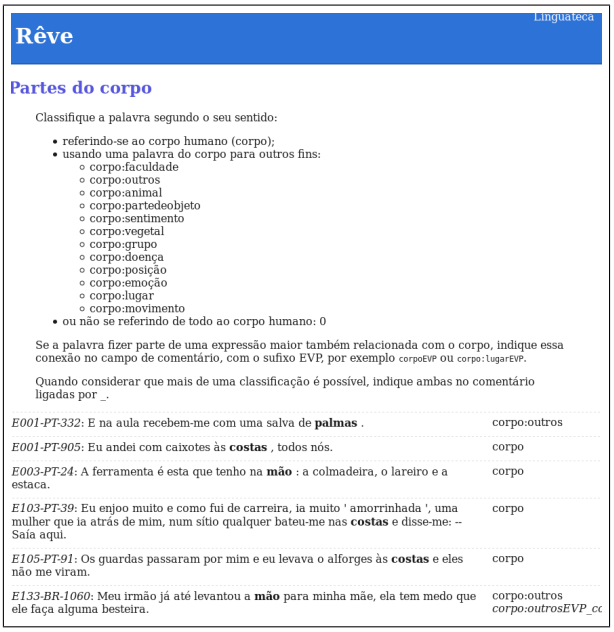



**Figura 6.** O Rêve aplicado à revisão de um conjunto de casos complicados da anotação do corpo humano a que estamos a proceder no Esqueleto.

## 6 Considerações finais

Parece-nos que a oferta destes serviços à comunidade do processamento computacional da língua portuguesa potencia o aproveitamento e a rentabilização do extenso e valioso material que já pomos à disponibilização de todos.

Muitas vezes, os utilizadores não estão conscientes das potencialidades ou mesmo da informação a que podem recorrer, por isso muitas vezes o mais difícil é tornar úteis ou mesmo inteligíveis recursos criados para estudar a língua. Como já discutido em [9], o artigo que representa a declaração de intenções sobre o que a Linguateca se propôs depois fazer, nem sempre é frutífera a colaboração entre (i) os compiladores de corpos, (ii) os usuários ou destinatários dos mesmos, e (iii) os desenvolvedores das ferramentas que lidam com corpos.

## Referências


1. Araújo, S., Almeida, J.J., Simões, A., Dias, I.: Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas. Linguamática 2(2), 71–74 (Junho 2010)
2. Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço AC/DC. In: The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009) (8-11 de Setembro 2009)
3. Evert, S., Team, T.O.D.: The IMS Open Corpus Workbench: CQP Query language Tutorial (2010), http://cwb.sourceforge.net/files/CQP_Tutorial.pdf
4. Freitas, C.: Esqueleto: Anotação das palavras do corpo humano (2013-2014), http://www.linguateca.pt/acesso/Esqueleto.pdf
5. Freitas, C., Santos, D., Gonçalo Oliveira, H., Quental, V.: VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In: Sarmento, S., Sardinha, T.B., Mottin, L.P., Ibaños, A.M.T. (eds.) Pesquisas e perspetivas em lingüística de corpus. pp. 199–232. Mercado de Letras, Campinas, Sao Paulo (2014)
6. Freitas, C., Santos, D., Sousa, R., Jansen, H., Mota, C.: Investigação do léxico do corpo humano e anotação semântica de corpus. In: ELC 2014 (2014)
7. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática 2(1), 77–93 (Maio 2010), nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009
8. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008), http://www.R-project.org, ISBN 3-900051-07-0
9. Santos, D.: Disponibilização de corpora de texto através da WWW. In: Marrafa, P., Mota, M.A. (eds.) Linguística Computacional: Investigação Fundamental e Aplicações. pp. 323–335. Colibri (1999)
10. Santos, D.: Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. In: Johannessen, J. (ed.) Language Variation Infrastructure: Papers on selected projects. pp. 113–128 (2011)
11. Santos, D.: Corpora at Linguateca: Vision and Roads Taken. In: Sardinha, T.B., Ferreira, T.S.B. (eds.) Working with Portuguese corpora. pp. 219–236. Bloomsbury (2014)
12. Santos, D.: Gramateca: corpus-based grammar of Portuguese. In: PROPOR2014. pp. 214–219 (Outubro 2014)
13. Santos, D.: PoNTE: apontando para corpos de aprendizes de tradução avançados. Linguamática 6(1) (Julho 2014)
14. Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000). pp. 205–210 (31 May-2 June 2000)
15. Santos, D., Marques, R.P.R., Freitas, C., Mota, C., Simões, A.: Comparando anotações na Gramateca. In: ELC 2014 (2014)
16. Simões, A.: Comparador: forma de auscultar corpos no AC/DC (Maio 2014), http://www.linguateca.pt/documentos/Comparador.pdf
17. Simões, A., Santos, D.: Ensinador: corpus-based Portuguese grammar exercises. Procesamiento del Lenguaje Natural 47, 301–309 (Setembro 2011)